

A REVIEW OF THE NEW AVIRIS DATA PROCESSING SYSTEM

Mikael Aronsson

Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109

1. INTRODUCTION

The processing of AVIRIS data - from VLDS¹ flight tape to delivered data products - has traditionally been performed in essentially the same way, from the beginning of the AVIRIS project up to and including the 1996 flight season. Starting with the 1997 flight season, a drastically different paradigm has been used for the processing of AVIRIS data. This change was made possible by the recent development of and related availability of affordable data storage devices.

2. DESCRIPTION OF THE PREVIOUS GENERATION DATA PROCESSING SYSTEM FOR AVIRIS

The first generation AVIRIS data processing system evolved, via incremental improvements from code originally developed for a VAX computer system, into a fairly complex set of software modules, executed on a network of Sun and Solbourne computers. The original code was written in FORTRAN and assembler, but by the time the migration to the UNIX-based systems took place, the code had been translated to C. The software modules were governed and driven by a Sybase relational database management system. The Sybase system was also used to store all the meta data (or, the data about the data) for all AVIRIS runs. The data from the runs themselves were stored in two magnetic tape-based forms. The first form used 10- or 12-bit data words² that were recorded in-flight on the VLDS flight tape. The second used 16-bit decommutated data words recorded on 4mm DAT tape, that was kept as a permanent archive of all AVIRIS runs.

Just like the AVIRIS instrument itself, the software modules were gradually improved over the years and had reached a significant level of complexity by the 1996 flight season, primarily caused by the desire to maintain backwards compatibility so that data from any flight season could be processed with the current version of the software modules. Here follows a brief description of the software modules' major functionalities as they appeared during the 1996 flight season (cf. figure 1). Before the first software module could be executed, information about the flight tape had to be entered in one of the Sybase tables. The first step of processing consisted of the VPS module, which operated on a complete flight tape, i.e., typically on a collection of 5 to 20 runs. VPS stands for VLDS PBN Scan, where PBN in turn stands for Principal Block Number. When data are recorded on the VLDS tape, they are laid down in consecutively numbered (and hence uniquely identified) swipes, each such swipe consisting of 64 kbytes of data. One 64 kbyte swipe is called a Principal Block. The major purpose of the VPS program module was to scan the VLDS tape in order to identify the Principal Block Number for the beginning and ending of each run, and also to estimate the size of each run in the process. This information was stored in a Sybase table.

Once the complete flight tape had been scanned by the VPS module, the remainder of the processing was performed on discrete runs or 6-scene parts of runs, should they be longer than 6 scenes. A scene is an artificial delimiter of each 512 lines (or major frames) of image data from a run. This delimiter had been

-
- 1 VLDS is the acronym for the Metrum Very Large Data Store storage system, which is the VHS-tape based recording system used to record data acquired by the AVIRIS instrument.
 - 2 AVIRIS data were recorded on the VLDS tapes as 10-bit raw data words in 1994 and earlier years. The word length was increased to 12 bits in 1995.

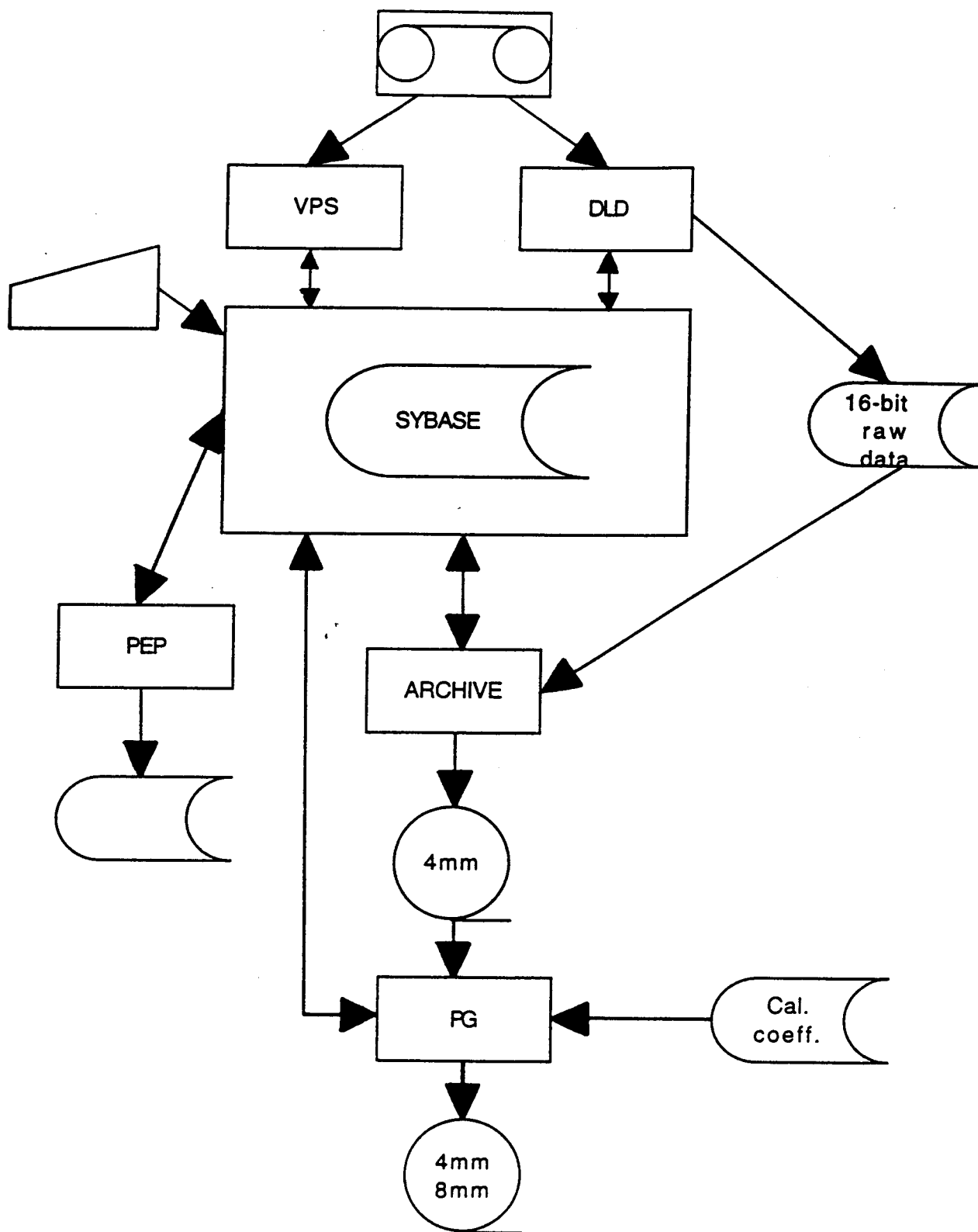


Figure 1. The first generation AVIRIS data processing system.

devised in the early days of AVIRIS, since this was the amount of data that would fit on one 9-track magnetic tape. The transition from 9-track tape to 4mm DAT tape allowed fitting larger amounts of data on each tape. The upper limit for what could fit on one tape was at this point not the delimiting factor any more. Instead, the delimiting factor was defined by the amount of AVIRIS data that could fit on any of the hard discs available in the AVIRIS Data Processing system (used for intermediate storage during processing of AVIRIS data), which was the origin of the 6-scene upper limit for AVIRIS datasets. The second step of processing, performed by the DLD module, would thus operate on datasets consisting of at most 6 scenes. DLD was an abbreviation of the word download, which describes the major action taken by this module, i.e., data was read from the VLDS tape and downloaded (or written) to a temporary disc file. Exactly which dataset would be downloaded was decided automatically by an algorithm that prodded one of the Sybase tables (even though a skilled operator could affect the order of downloading by manipulating the database). The raw 10- or 12-bit data words, as written on the VLDS tape, were converted to 16-bit integers as part of the process and more meta data was extracted and entered in Sybase tables. Since this was done in parallel with reading the data off of the VLDS tape and writing it to disc, the DLD module was rather slow and time consuming.

Once DLD had been completed, the data could be moved from the temporary disc file to a permanent archival storage device. This was done by the program module ARCHIVE and this program would write the AVIRIS data, on a scene-by-scene basis, to a 4mm DAT archive tape. Each archive tape would include a precal file, up to six scenes of science data, and a postcal file. In the early years there was only one copy of each archive tape made, but learning from experience, we eventually switched to making two copies of each tape. The ARCHIVE program module would also extract information to be used by the following program module and write this information to Sybase tables. ARCHIVE furthermore added information to yet another Sybase table in order to keep track of what dataset was archived on what tape.

The next step of processing was done by a program module called PEP, which was an acronym for Performance Evaluation Program. This program module would utilize the extracted information from the AVIRIS data in order to facilitate monitoring of the AVIRIS instrument's performance, as well as the performance of the VLDS recorder. The output from this module was automatically e-mailed to several individuals within the AVIRIS project and also stored in Sybase tables (for later perusal and trend analysis).

The final step of processing, called PG (for Product Generation), would be performed on request only. An investigator - or anyone else interested in a specific dataset (or run) - would submit a request, including information about what flight, run, and scene(s) were desired. This information, as well as a significant amount of additional information related to the request, would be entered into Sybase tables. The PG program module would then prod the tables, read the corresponding data off of the appropriate archive tape, calibrate the data, and write the resulting files to a PG tape. Depending on the investigator's request, the PG tape could be either a 4mm DAT tape or an 8mm EXABYTE tape and the data would be ordered and processed on a scene-by-scene basis, with a maximum of six scenes included per tape. In addition to the (one to six scenes of) science data, each PG tape would also include a number of ancillary files, e.g., pre- and postcalibration data, engineering and navigation data, calibration data, documentation, etc. The PG module would also produce a number of additional products, e.g., single-band images, spectral plots, etc., to be used in-house for quality control.

The processing of AVIRIS data was a slow and cumbersome process because all processing was governed by, and the software modules frequently interacted with, the Sybase relational database management system. Another reason for the slow processing speed was the multiple steps of tape input/output (tape i/o).

3. THE CURRENT GENERATION DATA PROCESSING SYSTEM

By the end of the 1996 flight season, large scale disc storage technology had become quite affordable, both as far as hard discs and virtual disc systems were concerned. The prospect of being able to access several years worth of AVIRIS data on-line, coupled with the desires to decrease the amount of tape i/o, to decrease or eliminate the reliance on the Sybase database system, to retire the old, increasingly complex software modules, and to speed up through-put, led to the decision to purchase several pieces of hardware and to design a new generation of software for processing of AVIRIS data. The hardware acquisitions included sixteen 9-GB and

eight 23-GB disc drives, first one and later a second 62-GB RAID-5³ system, and an AMASS⁴ system, originally designed to hold 2 TB of data, but recently upgraded to 3 TB. A set of new software modules was developed in tandem with these hardware acquisitions. We adopted something we refer to as the "KISS" principle for the development of the software, where "KISS" stands for "Keep It Simple, Stupid". This meant that, instead of designing large and complex program modules that would handle multiple tasks and be compatible no matter what year the data to be processed were from, smaller program modules, that were designed to handle only one year's data, were written (even though some modules do in reality not change from one year to another). Each module was designed to handle only a limited set of tasks, in order to keep the modules simple and easily maintainable. The new software does not interface with the Sybase database system and processing is hence not driven by the database. Instead, the new software system is governed by the operator running the software, i.e., he or she decides what program module to run and on what dataset. And, finally, the input to and output from the different program modules are kept on disc whenever possible, which significantly has speeded up the processing time. Here follows a brief description of the software modules of the new AVIRIS data processing system (cf. figure 2).

The first step of processing is performed by the VTOD program module, where VTOD stands for VLDS TO Disc. This program first scans the VLDS tape in order to find the filemarks identifying the end of each run. The data are then copied verbatim (i.e., as 12-bit raw data words) to disc, with each run making up a uniquely named file. The only exception from this rule occurs when a run is longer than approximately 7,400 major frames (or lines), which in accordance with the old way of thinking, equalled about 14.5 scenes. This, in turn, was approximately equal to a disc file size of 1.5 GB and the following paragraph will explain the rationale behind this limitation. VTOD normally operates on a complete flight tape.

The second software module is called EXP, which is short for EXPansion. The EXP module reads the raw 12-bit data, identifies the major frame sync words, and expands each 12-bit data word to a 16-bit data word. The result is written to disc and the resulting file is approximately 33% larger than the corresponding input file. This means that the maximum 12-bit raw data input file size of 1.5 GB is expanded to a 2 GB output file. Two GB is the largest file size the UNIX operating system can handle, hence the limitation of approximately 7,400 lines per file. The resulting output files are finally moved from regular hard disc to the AMASS system for permanent storage. This program module normally operates on all files from a flight tape as one batch job.

The third step of processing is done by the PE program module, where PE stands for Performance Evaluation. This module extracts information from the precal portion of the file output by EXP. The information extracted includes background noise and the difference between high (shutter open) and dark (shutter closed) signal for the four spectrometers, detector, spectrometer, scanner optics and on-board calibrator temperatures, and more. These values are written to a file and a number of key parameters are flagged if they do not meet predetermined threshold values, which could indicate a problem with the instrument. The file containing this information is saved for future trend analysis and it is also automatically e-mailed to certain key persons.

-
- 3 The RAID-5 is a hard disc system, consisting of eight 9-GB discs with accompanying proprietary VERITAS software. RAID is an acronym for Redundant Array of Inexpensive Disks, which means that the discs are set up so that part of the combined storage capacity is used for storing duplicate information about the data stored in the disc array. The duplicate data allows regeneration of the data in case of disc failure. The suffix 5 designates a specific type of RAID system, where the data is "striped" (or interleaved) across the discs in the array and the data redundancy is provided by the use of parity information. The complete set of discs appears to the user as one large disc.
 - 4 The AMASS Storage Management System from EMASS is a virtual disc system consisting of a hard disc cache and four DLT (Digital Linear Tape) drives with automatic loading/unloading of tapes. It is governed by an on-line AMASS database. Even though the bulk of the 3 TB of data are stored on DLT tape, data is automatically swapped between tape and the disc cache, giving the user the impression that all the data are available on-line (i.e., as if all the data were residing on hard disc).

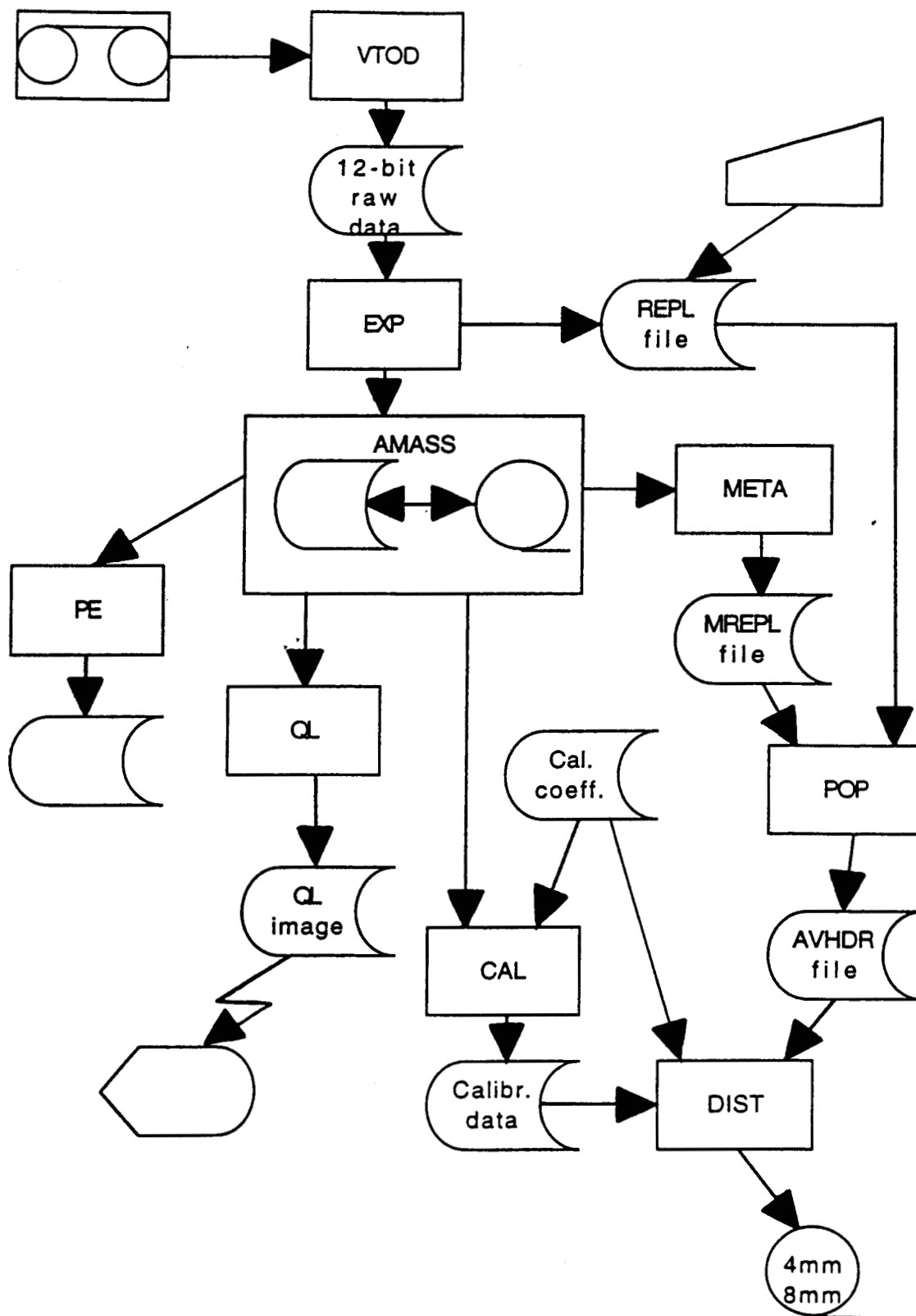


Figure 2. The second generation AVIRIS data processing system.

Next follows a manual step, where the operator enters information related to a particular run into an ASCII text file, referred to as the AVHDR (short for AVIRIS header) file. This is currently done via a text editor, such as emacs or vi, but will eventually be done via a GUI (= Graphical User's Interface). Among the information entered is Site name, Location, Investigator, Start and End Date, Time, Latitude and Longitude (nominal values, as given by the pilot), Comments about the run (from the pilot, the AVIRIS experiment coordinator and/or the operator), etc. Since this currently is a manual step, operation is limited to one file at a time.

The next software module is called META, and this module extracts meta data from each run. This meta data (or data about the data) includes Major Frame Counts (both total and by major frame type), Start and End Date, Time, Latitude, Longitude and Altitude (from the navigation and engineering data), as well as indications of missing lines, etc. META normally operates on all runs from a flight tape as one batch job.

The information manually entered by the operator and automatically extracted by the META software module is not entered directly into the actual AVHDR file. Instead, the information resides in intermediate files that are merged and copied to the AVHDR file by the next program module, called POP (which is short for POPulate, since this program populates the real AVHDR file with data). The POP module also extracts some of the information from the AVHDR file for generation of the Quicklook index file (see below).

Next follows the QL software module. This module generates a (subsampling) Quicklook (or QL) image of one band (normally band 34) of the science data in an AVIRIS run. The output is converted to JPEG format and the resulting file is moved to the AVIRIS FTP site, so that the QL image can be accessed by the investigators as well as any other interested party. The updated Quicklook index file (generated by POP, as described in the previous paragraph) is also moved to the FTP site.

At this point it is time to calibrate the data. This is done by a software module, appropriately named CAL, that reads the expanded raw data, extracts pre- and postcal data, as well as navigation, engineering, and dark signal data. CAL, furthermore, performs the calibration of the science data and also generates a browse image file and a number of quality control files (to be used in-house). All generated files are written to disc in preparation for the following step.

The final step of processing in the new software system is performed by a module named DIST (which is short for DISTribution). This module assembles the files produced by CAL (except for the in-house quality control files), the AVHDR file generated by the POP module, calibration constants files, and a documentation (or readme) file. All these files are placed in a special directory and the UNIX tar command, that creates a tape archive of the files, is issued. The resulting tar file is copied to a 4mm DAT or 8mm EXABYTE tape, which is sent to the investigator.

4. PERFORMANCE COMPARISON

It was estimated in the early days of AVIRIS' existence that the archiving of a completely full flight tape would take approximately 9 working days and that the retrieval processing (or product generation) would take an additional 60 working days (Reimer, et al., 1987). By 1992 the original processing system had been replaced by a Sun-based system and at that time the archiving of a full flight tape could be accomplished in 5 working days, while the product generation required approximately 5-10 working days (Hansen, et al., 1992). Through continued improvements over the years, the processing time had been further reduced by 1996, so that the archiving process could be accomplished in 3-4 working days and the product generation could be completed in another 5-7 working days, for a complete processing time of approximately 10 working days per flight (or VLDS) tape.

Once the new generation software and hardware system was completely in place (by the summer of 1997) we were able to process a completely full flight tape - all the way from VLDS tape to distributed data products - in the span of two working days, which represents a five-fold reduction in processing time.

5. FUTURE PLANS

Some of the software modules are currently used in their beta test versions. The process of replacing the beta modules with production versions has top priority in 1998. One of the beta modules is DIST, that currently can only produce tapes in tar format. It is anticipated that the conversion to the production version of this module will also include an upgrade that will allow us to produce distribution tapes that will have each discrete file written as a separate file on the tape (in a similar fashion to the PG tapes produced by the old software system) instead of bundling all files into one tar file.

As mentioned earlier, the manual step of entering the flight log information is scheduled for an upgrade to a GUI. In the long term we are even considering eliminating this step as part of the post-flight data processing done at JPL. We will instead have the Experiment Coordinator file this information electronically at the time of the flight.

At the present time, most of the software modules must be manually executed. In the long term this may get replaced with an automated system. We are looking into at least two different options, one being the usage of so-called cron jobs. A cron job is an automatic process that executes one (or several) software modules at a specific time each day. We could thus have a cron job that, e.g., at midnight each day identifies all new (and hence unprocessed) AVIRIS runs residing on the disc system and processes them. In fact, we are already experimenting with this option on a limited scale. Another option would utilize scripts that link the different software modules together, so that when one module, e.g., VTOD has successfully completed processing of a VLDS tape, the next module (which is EXP) would be automatically initiated, etc.

Besides processing all the data from the 1997 flight season, we have begun reprocessing the previous years' data with the new software. The plan is to continue this effort until all data back to 1992 have been reprocessed.

We are also planning a number of hardware upgrades. The computer used for processing of all the data will get upgraded CPUs, which should improve the processing throughput. Another upgrade that is planned for the coming year is the acquisition of a 0.5 TB hardware RAID-3⁵ disc subsystem. This disc system will allow us to store a whole year's worth of AVIRIS runs simultaneously on hard disc, thus enabling rapid comparisons, trend analysis, and other processing on a complete flight season.

6. ACKNOWLEDGEMENT

The work described in this paper was performed at the Jet Propulsion Laboratory under a contract with National Aeronautics and Space Administration.

7. REFERENCES

Hansen, E.G., Larson, S., Novack, H.I., Bennett, R., "AVIRIS ground data processing system", Summ. Third Annual JPL Airborne Geoscience Workshop, JPL Publication 92-14, Jet Propulsion Laboratory, Pasadena, CA, Vol. 1, 1992, pp. 80-82.

Reimer, J.H., Heyada, J.R., Carpenter, S.C., Deich, W.T.S., Lee, M., "Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) ground data-processing system", Imaging Spectroscopy II, Proceedings of SPIE, Volume 834, 1987, pp. 79-90.

5 The RAID-3 system is similar to the RAID-5 system, with the major difference being exactly how the data stripes and parity information is laid out across the discs.